ISSN 2464-4617(print)
ISSN 2464-4625(CD)
Computer Science Research Notes
CSRN 2803
Poster's Proceedings
http://www.WSCG.eu

A new model driven architecture for deep learning-based multimodal lifelog retrieval

Fatma Ben Abdallah
Regim-LAB
REsearch Groups in Intelligent Machines,
University of Sfax,
National Engineering School of Sfax (ENIS)
3038, Sfax, Tunisia
ben.abdallah.fatma@ieee.org

Ghada Feki
Regim-LAB
REsearch Groups in Intelligent Machines,
University of Sfax,
National Engineering School of Sfax (ENIS)
3038, Sfax, Tunisia
ghada.feki@ieee.org

Anis Ben Ammar
Regim-LAB
REsearch Groups in Intelligent Machines,
University of Sfax,
National Engineering School of Sfax (ENIS)
3038, Sfax, Tunisia
anis.ben.ammar@ieee.org

Chokri Ben Amar
Regim-LAB
REsearch Groups in Intelligent Machines,
University of Sfax,
National Engineering School of Sfax (ENIS)
3038, Sfax, Tunisia
chokri.benamar@ieee.org

ABSTRACT

Nowadays, taking photos and recording our life are daily task for the majority of people. The recorded information helped to build several applications like the self-monitoring of activities, memory assistance and long-term assisted living. This trend, called lifelogging, interests a lot of research communities such as computer vision, machine learning, human-computer interaction, pervasive computing and multimedia. Great effort have been made in the acquisition and the storage of captured data but there are still challenges in managing, analyzing, indexing, retrieving, summarizing and visualizing these captured data. In this work, we present a new model driven architecture for deep learning-based multimodal lifelog retrieval, summarization and visualization. Our proposed approach is based on different models integrated in an architecture established on four phases. Based on Convolutional Neural Network, the first phase consists of data preprocessing for discarding noisy images. In a second step, we extract several features to enhance the data description. Then, we generate a semantic segmentation to limit the search area in order to better control the runtime and the complexity. The second phase consist in analyzing the query. The third phase which based on Relational Network aims at retrieving the data matching the query. The final phase treat the diversity-based summarization with k-means which offers, to lifelogger, a key-frame concept and context selection-based visualization.

Keywords

Lifelogging, Multimodality, Retrieval, Summarization, Visualization, Convolutional Neural Network , Relational Network.

1 INTRODUCTION

Recently, we have witnessed the emergence of user-centric approaches in multimedia retrieval [Fek16] [Fak16] [Bouh17] [Gue11] [Wal10]. In fact, personalizing the search is the main objective in several on-going research domains like egocentric vision, self-tracking, quantified-self and personal data which are more commonly known for the last decade as lifelogging. Indeed, lifelog consists of acquiring data via cameras and sensors and storing this data to form a personal archive [Gur14]. Since the dawn of time, men have always tried to leave traces of their activities and their daily lives. Prehistoric men painted frescoes in the caves. Later, men recorded their thoughts, their moods and their days in diaries. Nowadays, a lot of wearable cameras have been created to facilitate the automatic capture of images and videos of daily life. In this work, we focused on data captured with photographic camera commonly called visual lifelogs, because with this kind of camera we can acquired over long periods of time which would not be possible with videos cameras. If lifelog's primary goal is to build a personal archive that extends over years, the best way to do this is to use images instead of videos. Lifelogging is characterized by the huge amount of personal data generated by the lifelogger. This data does not contain neither annotations nor semantic descriptions. An effort has been done these last years to construct lifelog datasets which contains tens of thousands of images. Considering this huge amount of personal data created, there is a need for systems that can automatically analyse the data in order to understand, classify, summarize and also query to retrieve the information the user may need. Another

https://doi.org/10.24132/CSRN.2018.2803.2
ISBN 978-80-86943-42-8

specificity concerns the images, they are captured automatically at regular intervals without knowing on what focus lifelogger's attention. Sometimes, images are noisy, error-prone, blurry, distorted or contains useless information like sky or walls. Operate a preprocessing is necessary to eliminate noisy information. To make this personal data usable, several personal lifelogging applications (PLA) have been realized [Gur14]. These PLA cover different life aspect of lifelogger, self-monitoring of activities (sport, dietary, sleeping, smoking cessation), memory assistance (help people with memory loss such as with Alzheimer's disease or dementia) and long-term assisted living (to prevent older adults from potential alarm situations).Despite the advances in the visual lifelogs capture and storage, relevant data mining from this considerable amount of multimodal data remains unresolved problem. To the best of our knowledge, three major issues are present in lifelogging. First, the multimodality is generally not addressed [Dan17]. Second, a complete system combining annotation, retrieval, summarization and visualization is not proposed yet. Third, the architecture of existing systems are not based on any model.

In this paper, we present a new model driven architecture for deep learning-based multimodal retrieval, summarization and visualization. Our proposed architecture consists of four phases. The first phase process begin with preprocessing the lifelog images using CNN. Then, an extraction feature with enhancement is operate relying on several CNN pretrained on Imagenet. After that, a semantic segmentation using Global Convolutional Network (GCN) limit the search area in order to better control the runtime and the complexity. The second phase, based on Relational Network (RN), consist in retrieve moments according to the user's query. The third phase summarize the output of retrieval based on diversity using convolutional k-means. The final phase gives the summary's visualization based on different concepts and contexts. The remainder of this paper is divided into five sections. In section 2, we present recent related works in the retrieval, summarization and visualization context and discuss on-going challenges in lifelogging. In section 3, we describe our model driven approach which is based on several conceptual model. Section 4 details the four phases of our new architecture which based on several deep learning method for multimedia lifelog retrieval and summarization. The section 5 provides some concluding remarks and suggests future works.

2 OVERVIEW OF CHALLENGES IN LIFELOGGING

For several years, proposing systems which are able to extract relevant information from lifelog data has been the interest of various researchers. This necessity is linked to the exponential growth of the data recorded by the various cameras and sensors.

Lifelogging has become a full-fledged task at international conferences and a special attention is paid to it. We are witnessed the creation of tasks and benchmarks in international conferences dedicated to lifelogging such as NTCIR [Gur17] and IMAGECLEFLifelog Task [Dan17] which gave rise to the construction of test collection for lifelog research. In the following, we will present an overview on retrieval, summarization and visualization lifelog approaches which, most of them, were proposed on this benchmarks.

2.1 Deep Learning based approaches of retrieval

The central theme of the retrieval is the study of models and systems of interaction between human users and corpus of digital documents in order to satisfy their information needs. In the context of lifelogging, information retrieval consists in finding an event, an object, a person, a place or an action in a huge personal archive. Our study focused on works based on deep learning in the context of egocentric image retrieval. We found that the majority of the studied works relied on pretrained CNN.

Authors in [Rey16] design a system based on Bag-of-Words framework able to help users to find their personal objects once they have forgotten or lost. The classification of the relevant candidates and the discarded ones is achieved by thresholding. In [Oli16b], authors proposes a text-based search engine approach on certainty score tf-idf for egocentric images retrieval based on CAFFENET[1] that takes advantages of the inverted index approach. In [Oli17], authors create an automatic method based on LSDA[2] framework for semantic indexing and retrieval of wearer activities in egocentric images based on integrating heterogenous information from images and metadata. They used the retrieval engine indexation based on certainty score developed in [Oli16b]. Authors in [Oli16a] design an interactive systems which employed a semantic content tagging mechanism based on the retrieval system open source LUCENE[3] image retrieval engine. Authors in [Xia16] based their work on feature expansion using Wordnet and a manually performed query expansion by an expert. [Saf16] used three Deep Convolutional Neural Network models (Alexnet, Googlnet and VGGnet) learned on the IMAGENET[4] corpus and Multiple SVM approach learned on the TRECVID2013[5] data using the CAFFE framework. Authors in [Lin16] proposed a textual approach based on word to vector model. They

[1] http://caffe.berkeleyvision.org/
[2] http://lsda.berkeleyvision.org/
[3] http://www.lire-project.net/
[4] http://www.image-net.org/
[5] https://www-nlpir.nist.gov/projects/trecvid/trecvid.data.html#tv13

ISBN 978-80-86943-42-8

use a word distance between the provided CAFFE concepts and the keywords from the quey. [Zhou17] used the human-in-the-loop methods to match between query and user needs.

By analyzing these works more closely, there are some limitations which we detail in the following. In [Rey16], bag of words ignores the context of words : it does not take into account semantic meaning and ordering. In [Oli16b], the authors used EDUB dataset 2015 to evaluate the approach. The dataset contain only 4192 images. With a bigger dataset, there will may be a problem of performance with the inverted index. That's why they used NTCIR-12 lifelog dataset in [Oli17]. Concerning [Oli16a], we have doubts about system adaptation to a change in the order of demand's magnitude, particularly in maintaining functionality and performance in high demand. [Xia16] not achieve promising result, they failed to provide good retrieval effectiveness as they said in their article. Authors in [Saf16] use temporal indexing which is fuzzy and culturally dependent. They also link the terms of the topic manually to the set of IMAGENET and TRECVID and generate manually the queries from the topics. In [Lin16], authors faces difficulties to construct the relations of topic question keywords and CAFFE concept words. [Zhou17] exploit only the information's provided which consist in the description of the semantic locations and physical activities.

2.2 Clustering based summarization

The summarization process aims to produce a concise version of one or more digital documents containing only the most important information, possibly responding to a user need. Ideally, information retrieval should use the abstract to present synthetic results requiring minimal time to be appreciated by the user. Summarization deals with unsupervised classification since we do not know in advance the number of classes which represent the image returned by the system. Two methods were used in the majority of the works presented below : hierarchical clustering and k-means.

In [Mol16], they filter uninformative images by analyzing their ratio edges and describes the images using the available CNN models for objects and places with egocentric-driven augmentation. Then, they cluster into episode using k-means approach. [Bol15] based their work on frames characterization by the means of the pretrained CAFFENET convnet, which will be segmented based on unsupervised hierarchical agglomerative clustering. Then to choose the best photo to represent the event , they select the most visually similar with the rest of the photos in the same cluster based on random walk and minimum distance. [Lid15] remove non-informative images by a new CNN-based filter. Then, images are ranked by relevance to ensure semantic diversity : relevance ranking was obtained by integrating techniques for saliency detection, object recognition and face detection. Moreover, re-rank is applied by enforcing diversity among the chosen subset of pictures. Finally, they define the priorities of the different relevance terms based on Mean Sum of Maximal Similarities. [Dog17] analyze the output of concept detector provided by the organizers and selecting for each image the most probable concepts. Then, they perform image clustering based on the histogram of oriented gradients (HOG) and stopped the hierarchical clustering algorithm when 30 clusters were formed. After that, they use WU-Palmer similarity score using Wordnet to calculate similarity be-tween each image from the cluster and the topic description. Finally, they sorted the clusters in descending order based on the mean value of the similarity scores of the images that it contained. [Mol17] choose to begin with preprocessing techniques to filter out uninformative images. Then, they rank the remaining images according to how well they match the given query. After that, they cluster the top ranked images into a series of events. Finally, they select images in interactive manner according to distance to the cluster for k-means clustering or relevance score for hierarchical trees. [Zhou17] use hierarchical clustering and select the top image that close to center for summarization. By looking at these works more attentively, we notice that [Mol16] and [Bol15] did not take into account diversity in the ranking. For [Mol16], the evaluated dataset is not big and representative enough and the pretrained models that they use do not properly classify egocentric images content. [Bol15] did not use semantic information like object, people or actions and rely solely on low-level features. [Lid15] did not take into account spatial and temporal information. [Dog17] provide results not satisfactory due to the lack of correlation between the concept output by the CAFFE concept detector and the meaning of textual descriptions of the topics. Besides, temporal information has not been used. They, relied solely on the information provided by the organizers and no additional annotations or external data have been used. For [Mol17], different tasks require different summarization method which may not be completely consistent when changing the lifelog input.

2.3 Multimodal Visualization

Managing, searching and browsing a large amount of lifelog images through an interface has been the subject of several research. [Lee08] was the first to create a web interface for browse, search, annotate or save for future reference Sensecam photos. [Oli16b] proposed a web based prototype too. [Oli16a] realize a heat map which sort the images by the time and highlight those that express the context of the moment. [Hwa13] propose a mobile life browser, called MylifeBrowser, which visualizes and searches the lifelog data from mo-

 ISBN 978-80-86943-42-8

bile device. [Lar13] propose quantified self (QS) Spiral an interactive visualization technique that aims to capture the quantified self-data and let the user explore those recurring patterns. [Hop13] present different visualization techniques like Comic-book style Visual diary inspired by the squarified treemap pattern, timeline, master detailed having the appearance of a thumbnail gallery, a social interaction radar graph and a activity yearly calendar. [Dua17] investigate in virtual reality by realizing a virtual reality lifelog prototype.
The works presented in the previous subsection summarization did not deal with visualization, they only display the result at screen. They do not offer a way to visualize a relevant images selection to a specific search. Indeed, generally the images captured using wearable camera can be consulted via the application on computer or on mobile phone. In this kind of application, images are sorted according to a timeline per day or where every picture was taken based on GPS sensor.
At the end of this overview, we can say that no system propose to combine annotation, retrieval, summarization and visualization given the colossal difficulty that this induces, the systems focus only on one or two of this phases but not on the all. Also, we have noticed that many works proposed systems or frameworks which are not based on any model. Furthermore, deep learning is used only for extraction feature. The majority of the proposed works relied solely on the information provided by the organizers and no additional annotations or external data have been used. In all the above mentioned approaches, performance can be improved when the CNN is retrained on images that are more related to the retrieval dataset according to [Bab14]. The majority of them trained the images on IMAGENET. They have resorted to manual annotation to fill this information gap. We focus on deep learning-based approaches since it achieved promising results compared to classical ones as they mentioned in [Kri12] and [Gar17].

3 NEW ARCHITECTURE FOR DEEP LEARNING-BASED MULTIMODAL LIFELOG RETRIEVAL, SUMMARIZATION AND VISUALIZATION

In the following, we detail the novel architecture for deep learning-based multimodal lifelog retrieval, summarization and visualization. We first describe the general architecture, then we detail every step.

3.1 Model driven

Existing works have focused on one problem at time : they treated either retrieval, summarization or visualization. But none of them, except [Zhou17] who combined retrieval and summarization, has tried to create a whole chain of processing starting from the data's preprocessing until the visualization through the retrieval and the summarization. To be able to integrate these different phases, we felt that it was essential to be based on a model. The model is an essential condition for realization of a solid architecture as well as a good understanding of the system to be developed. A model can be considered as an abstraction of a system in the form of a set of facts. We choose to use model driven approach to realize our system. To do this, we will use UML language which is considered as a platform-independent modeling language and used in model-driven architecture introduced by the Object Management Group. UML model systems according to different points of view, static and dynamic. The static structure allows to model a system using objects, attributes, operations, and relationships. We described this static structure through the use case and the class diagram. We chose to model the dynamic behavior and the interactions between objects through the activity diagram. To elaborate this diagrams, we rely on the 5R's described in [Sel10] : recollecting, reminiscing, retrieving information, reflecting and remembering intentions. More details are shown in Fig 1.

3.1.1 Metamodel for Lifelogger

The lifelogger is the main actor of our system. He's who will create the database by capturing images using a device such as a sensor, camera or smartphone. Each image will have as unique identifier the date on which it was taken. When the lifelogger wants to search for a moment, he may be able to search for an object, a person, an event, a place, an emotion or an activity daily living (ADL) context and concept-based composed of individual actions. The result should be summarized before visualization. The fig.2 shows the map for modelling lifelogger needs.

3.1.2 Activity diagram

The activity diagram is a dynamic UML diagram describing the sequential activities and parallel systems. They allow to represent graphically the behavior of a method. The fig.3.a described the sequencing of a captured image and a search for a moment through the system. Following the preprocessing, the image can be deleted if it is uninformative. If the image is blurred, the system try to unblur it. If the image contain homogenous color this will mean that the image is uninformative and it will be deleted. Then, the system extract image feature and try to detect concepts. If the system find a new concept, it then make automatic annotation to the image. In both case, when finding or no a new concept, the next step is to realize a semantic segmentation. After that, when the lifelogger submit a query to find a specific moment, the system active the retrieval phase. Afterward, the system rank the result

 ISBN 978-80-86943-42-8

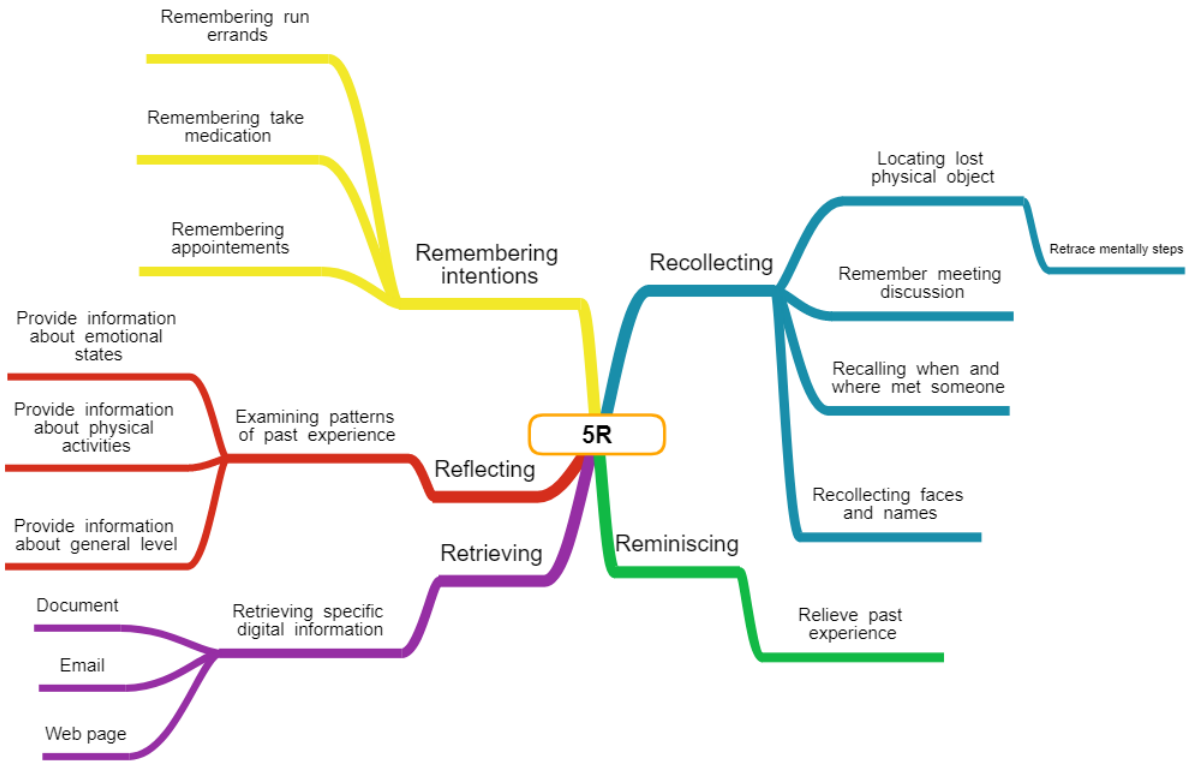


Figure 1: 5Rs mind map

of the retrieval and summarize it. The outcome is then visualize by the lifelogger.

3.2 Proposed architecture

Deep learning (DL) uses supervised learning based on digital artificial neural networks to allows a program, for example, to recognize the content of an image [Fak17] or to understand spoken language like Siri, Cortana and Google Now. With traditional methods, the machine simply compares the pixels. DL allows learning on more abstract characteristics than pixel values, which it will auto-construct. The knowledge on the classification of images contained in such network can be exploited in two ways: as an automatic extractor features, materialized by the CNN code and as a Fine Tuning to deal with the new classification problem. Given the power of DL in recognition and image processing [Boug14], we chose to use this method in our architecture.

The input of our architecture is a lifelog dataset which contains images, images concepts, extra data and queries. These multimodal and heterogeneous data was recorded from several acquisition modalities and came in different formats. The first one contain three steps: preprocessing, image feature extraction and semantic segmentation. This phase aims to delete uninformative image, to unblur blurred image by using pre-trained CNN as a feature extractor and to group images in homogeneous segment. We notice that performance can be improved when the CNN is retrained on images that are more related to the retrieval dataset. We will then also work on training the CNN on annotated lifelog dataset to improve performance and relevance. In the second phase, we use LSTM encoder to realize query processing. Since we will use the Relational Network to solve the retrieval problem in the third phase, it is better to perform this encoding. After ranking the result returned by the retrieval step using convolutional k-means, our architecture summarize the result based on diversity using keyframe selection. Finally, the result is shown to the lifelogger in a personalized way based on concepts and contexts. In the following, we will detailed every step includes in the architecture shown by the fig.3.b.

3.2.1 Automatic annotation enhancement

The images captured using wearable camera have a particularity. Indeed, these images are captured automatically at regular intervals (every 30 seconds) which causes repetitive shooting. Moreover, these images are sometimes taken in bad conditions like bad lighting or bad framing which give blurry images. To judge the quality of an image, we choose a CNN for no-reference image quality assessment [Bos16] applied to blurriness

ISBN 978-80-86943-42-8

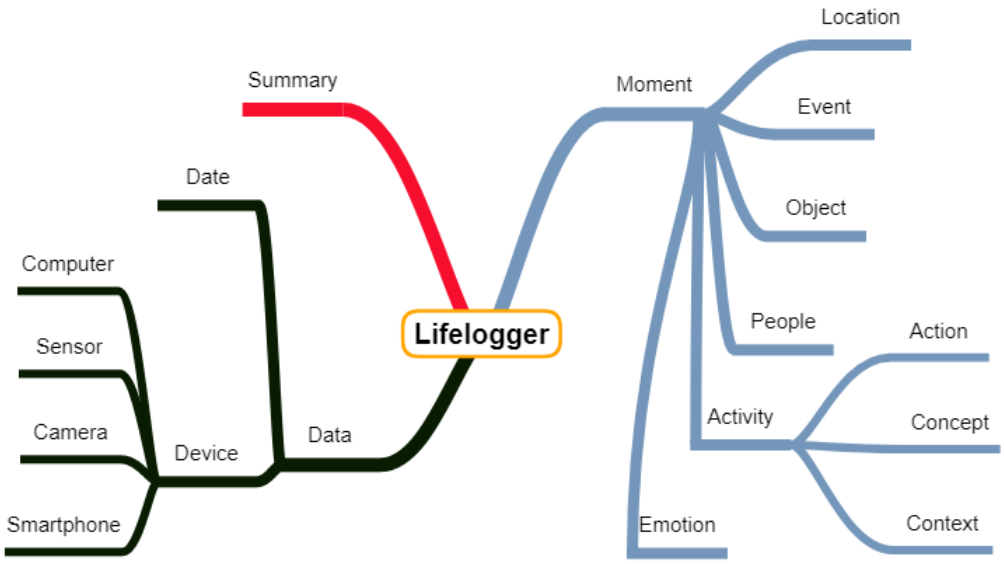
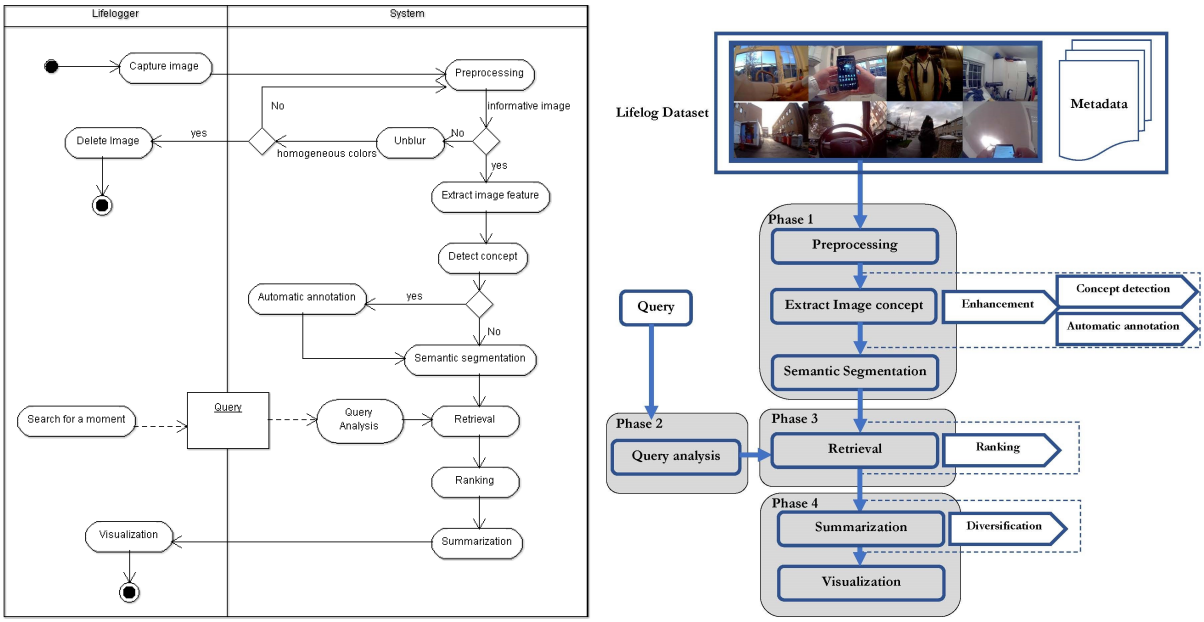


Figure 2: Lifelogger needs map



(a) Activity diagram of lifelog retrieval, summarization and visualization system

(b) Proposed architecture

Figure 3: General architecture

and colour diversity. If the image is blurred, we try to unblur it otherwise if the image contains homogeneous colour this means that the image is uninformative and should be deleted because it may be contain a sky, a wall or a floor.

The image is no longer considered as a matrix of colours' pixels, but as a carrier of a semantics encompassing several concepts. Convolutional neural networks have made considerable progress in the analysis of images especially on large dataset. To extract image feature, we rely on several CNN trained on Imagenet. In this step, we proceed to enhancement through a combination of concept detection by using the LSDA object detector which transfer classifiers for categories into detectors and automatic annotation if the discovered concept in the image do not appear in the list of concepts already related to the image in the lifelog dataset.

The objective of the semantic segmentation is the association of each object of the image, a label among a set of predefined classes (human activities, food, computer activity, heart rate ...). The final goal is to predict a mask of segmentation that indicates the category of each object. Pixels are classified based on

ISBN 978-80-86943-42-8

characteristics extracted in the image feature extraction step. By forming segments, we reduce the number of images to be processed in the retrieval phase. For that purpose, we will use a global convolutional network described in [Pen17] based on scalable ontology driven framework for hierarchical concept detection.

3.2.2 Query analysis

In the Lifelog Semantic Access Task (LSAT) of NTCIR-12 and in the lifelog retrieval task (LRT) of IMAGECLEF2017, the query is a set of topics representing lifelogger's information needs. Therefore, it is necessary to filter this query in order to extract the key concepts based on the following axes: object, location, event, ADL, people or emotion. Considering that we will use the Relational Network [San17] in the retrieval phase, the query should go through LSTM which it is capable of learning long-term dependencies.

3.2.3 Image retrieval

Information retrieval models define a representation of documents and the queries as well as a correspondence function which makes it possible to calculate similarities between documents and queries and to rank the results. Whatever the research model and the calculations are purely numerical and rely essentially on the frequency of words and analysis of their distribution, the search for semantic information seeks to go beyond this approach by injecting knowledge [Fek15]. For that purpose, we choose to adapt Relational Network (RN) by using Neural Tensor Network instead of Multi-layer Perceptron. To use RN, we need to build a reasoning lifelog dataset, which contain images and questions that test logical reasoning to enable detailed analysis of visual reasoning. We will then rank the result using Support Vector Machine.

3.2.4 Summarization and Visualization

The information's relevance returned to the lifelogger is an important aspect of the information retrieval but we should also take into account the diversity. Images sorted by relevance may be similar and redundant. It is not interesting to rank all similar image of the same one, even if it may be the most relevant. Instead, the top results should be different and complementary. Once the images are ranked, we use a k-means clustering based on average distance to apply diversification [Fek17] [Fek14] [Ksi13] on the result that will be visualized. The top n images from each cluster will be selected where n is a limit fixed by the user.

According to the user-study done by [Cho16], visualization should be insightful, intuitive, interactive, impressive and immersive. Also, the studied lifeloggers prefer a visualization based on the most frequently visited locations in visually appealing and informative interface that another based on temporal clustering, tempo-spatial clustering or tempo-spatial-visual clustering. We offer flexible visualization's interface in order to respond to various user needs. We also investigate in describing as the frequency and spending time for activities of daily living concepts (exp. : commuting , travelling, preparing meal, ...) and total time for contexts (exp. : in an office environment, in a home, in an open space, ...).

4 EXPERIMENTS

To evaluate the performance of our method, we employed ImageCLEFlifelog2017 which is based on the data available for the NTCIR12Lifelog task. ImageCLEFlifelog2017 was gathered by 3 lifeloggers during one month giving 79 days of data. It contains 88 124 images acquired using OMG Autographer wearable camera, XML description of semantic locations and the physical activities of each lifelogger at one minute. The output of the CAFFE CNN-based visual concept detector was included in the test collection.

We realized a preliminary experimentation for the automatic annotation enhancement using Matlab neural network toolbox. To extract image concept, we rely on four CNN : VGG-19, Resnet-50, Resnet-101 and InceptionV3 trained on Imagenet. The concept detection is based on CNNs choice which was guided by an analysis of deep neural network models for practical applications [Can16]. The analysis compare AlexNet, BN AlexNet, BN NIN, ENet, GoogleNet, VGG-16, VGG-19, Resnet-18, Resnet-34, Resnet-50, Resnet-101, Resnet-152, InceptionV3 and Inception V4 in term of accuracy, parameters, memory footprint, power consumption, operations count and inference time. According to the results of the analysis, we focus on the three top CNN that have provided the best result in accuracy.

An example of concept detection is shown in fig. 4. We split the image because the input of the VGG19, Resnet50 and Resnet101 CNN should be an image with frame size of 224x224. For InceptionV3, the frame size should be 299x299.

The table 1 details the relevance of each detected concept for the corresponding frame. We use several CNNs with different numbers of layers. The CNN layers consist of convolutional layers, pooling layers, fully connected layers and normalization layers. We notice that InceptionV3, which contains 316 layers, generates 5 relevant concepts. In fact, the concept "Coffee Mug" is detected thanks to the wider frame size compared to the frame size of the Resnet-50, Resnet-101 and VGG-19. This object is not detected since it is located on the border by the other CNN. For the architecture which have the same frame size, we notice that Resnet-101 which contains 347 layers, generates 2 out of 20 relevant concepts.

 ISBN 978-80-86943-42-8

(a) With VGG19, Resnet50 and Resnet101 (b) With Inception V3

Figure 4: Example of concept detection

Frame	Resnet50	R	I	Resnet101	R	I	VGG19	R	I	InceptionV3	R	I
1	Groom		√	Groom		√	Mosquito net		√	Coffee Mug	√	
2	Wall clock		√	Book jacket		√	Ipod		√	Loud Speaker		√
3	Notebook		√	Ipod		√	Notebook		√	Chime		√
4	Refrigerator		√	Wash basin		√	Chime		√	Photocopier		√
5	Photocopier		√	Refrigerator		√	Payphone		√	Medecine chest		√
6	Beer bottle	√		Laptop		√	Rubber eraser		√	Computer Keybord		√
7	Cash machine		√	Medecine chest	√		Ipod		√	Banjo		√
8	Shoji		√	Cash machine		√	Notebook		√	Casette player		√
9	Wash basin		√	Dish washer		√	Window shade		√	Toaster	√	
10	Photocopier		√	Stove		√	Space bar		√	Coffepot	√	
11	Doormat		√	Photocopier		√	Laptop		√	Frying pan	√	
12	Tub		√	Radiator		√	Prison		√	Watter bottle	√	
13	Carton		√	Tub		√	Banjo		√			
14	Wash Basin		√	Tub		√	Prison		√			
15	Cassette		√	Wash basin		√	Tape player		√			
16	Bannister		√	CD player		√	Bannister		√			
17	Coffepot	√		Planetarium		√	Coffepot	√				
18	Coffepot	√		Coffepot	√		Vaccum		√			
19	Nipple		√	Thimble		√	Oil filter	√				
20	Water bottle	√		Nipple		√	Pill bottle	√				

Table 1: Top20 concepts detected with different CNNs(**R**:Relevant,**I**:Irrelevant)

Although, it contains 177 layers, Resnet-50 generates 4 out of 20 relevant concepts. The reason for this difference is due to the single-crop error rates : models perform better when using more than one crop at test-time. Similarly, VGG-19 generates 3 out of 20 relevant concepts with 47 layers.

5 CONCLUSION

In this paper, we have addressed the creation of a novel model driven architecture for deep learning-based multimodal lifelog retrieval, summarization and visualization. The architecture consist of four phases integrating several conceptual models. The first phase process begin with preprocessing the lifelog images using CNN.

ISBN 978-80-86943-42-8

Then, an extraction feature with enhancement is operate relying on several CNN trained on Imagenet. After that, a semantic segmentation, using GCN, limit the search area in order to better control the runtime and the complexity. The second phase, based on RN, consist in retrieve moments according to the user's query. The third phase summarize the output of retrieval based on diversity using convolutional k-means. The final phase gives the summary's visualization based on different concepts and contexts. As future works, we are making implementation of the several phases of our system and we aim to validate our architecture by participating at IMAGECLEF 2018 Lifelog Task.

6 ACKNOWLEDGMENTS

The research leading to these results has received funding from the Ministry of Higher Education and Scientific Research of Tunisia under the grant agreement number LR11ES48.

7 REFERENCES

[Bab14] Babenko A., Slesarev A., Chigorin A., Lempitsky V. Neural Codes for Image Retrieval. In: Fleet D., Pajdla T., Schiele B., Tuytelaars T. (eds) Computer Vision - ECCV 2014. Lecture Notes in Computer Science, vol 8689. Springer, Cham.

[Bol15] Bolanos M., Mestre R., Talavera E., Giro X., Radeva P., Visual summary of egocentric photostreams by representative keyframes, ICME Workshops, pp. 1-6, 2015.

[Bos16] Bosse S., Maniry D., Wiegand T., Samek W.:A deep neural network for image quality assessment. ICIP 2016:pp 3773-3777, 2016.

[Boug14] Boughrara H., Chtourou, M., Ben Amar C., & Chen, L., Facial expression recognition based on a mlp neural network using constructive training algorithm, Multimedia Tools and Applications, pp. 709-731, 2014.

[Bouh17] Bouhlel N.; Feki G.; Ben Ammar A. & Ben Amar C., A Hypergraph-Based Reranking Model for Retrieving Diverse Social Images , Computer Analysis of Images and Patterns, Springer International Publishing, pp. 279-291, 2017.

[Can16] Canziani A., Paszke A. & Culurciello, E., An Analysis of Deep Neural Network Models for Practical Applications, CoRR, 2016 , abs/1605.07678.

[Cho16] Chowdhury S., Ferdous M. S. & Jose J. M., A user-study examining visualization of lifelogs, 2016 14th International Workshop on Content-Based Multimedia Indexing (CBMI), pp. 1-6, 2016.

[Dan17] Dang-Nguyen D.-T., Piras L., Riegler M., Boato G., Zhou L. & Gurrin C., Overview of ImageCLEFlifelog 2017: Lifelog Retrieval and Summarization, CLEF2017 Working Notes, CEUR-WS, volume 1866, 2017.

[Dog17] Dogariu, M. & Ionescu, B., A Textual Filtering of HOG-Based Hierarchical Clustering of Lifelog Data, Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017.

[Dua17] Duane A., Gurrin C., Investigating virtual reality as a tool for visual lifelog exploration(2017).

[Fak16] Fakhfakh R., Feki G., Ammar, A. B. & Amar C. B. Personalizing information retrieval: A new model for user preferences elicitation, IEEE International Conference on Systems, Man and Cybernetics (SMC), 2016.

[Fak17] Fakhfakh R., Ben Ammar A. & Ben Amar C., Deep Learning-Based Recommendation: Current Issues and Challenges, International Journal of Advanced Computer Science and Applications(IJACSA), 8(12), 2017.

[Fek14] Feki G., Ben Ammar A. & Ben Amar C., Adaptive Semantic Construction for Diversity-based Image Retrieval, Proceedings of the International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - Volume 1, SCITEPRESS - Science and Technology Publications, LDA, pp. 444-449, 2014.

[Fek15] Feki G., Fakhfakh R., Ammar A. B. & Amar C. B., Knowledge structures: Which one to use for the query disambiguation? 2015 15th International Conference on Intelligent Systems Design and Applications (ISDA), pp. 499-504, 2015.

[Fek16] Feki G., Feki, G., Ammar A. B. & Amar C. B. Query Disambiguation: User-centric Approach, Journal of Information Assurance Security, Vol. 11 Issue 3, pp. 144-156, 2016.

[Fek17] Feki G., Ben Ammar A. & Ben Amar C., Towards diverse visual suggestions on Flickr, Proceedings Volume 10341, Ninth International Conference on Machine Vision (ICMV 2016), 2017.

[Gar17] Garcia-Garcia A., Orts-Escolano S., Oprea S., Villena-Martinez V., Garcia-Rodriguez J. A review on deep learning techniques applied to semantic segmentation, arXiv:1704.06857, 2017.

[Gue11] Guedri B., Zaied M. & Amar C. B., Indexing and images retrieval by content, International Conference on High Performance Computing Simulation, ,pp. 369-375, 2011.

[Gur14] Gurrin C., Smeaton A. F. & Doherty A. R., LifeLogging: Personal Big Data, Foundations and Trends in Information Retrieval, pp 1-125, 2014.

 ISBN 978-80-86943-42-8

[Gur17] Gurrin C., Joho H., Hopfgartner F., Zhou L., Gupta R., Albatal R. & Dang Nguyen D. T., Overview of NTCIR-13 Lifelog-2 Task, Thirteenth NTCIR conference (NTCIR-13), Tokyo, Japan, 5-8 Dec 2017.

[Hop13] Hopfgartner F., Yang Y., Zhou L., Gurrin, C. User Interaction Templates for the Design of Lifelogging Systems, In Semantic Models for Adaptive Interactive Systems, pp. 187-204, 2013.

[Hwa13] Hwang K.S., Cho S.B., A Lifelog browser for visualization and search of mobile everyday-life, Mobile Information Systems, 10, pp. 243-258, 2013.

[Kri12] Krizhevsky A., Sutskever I., Hinton G.E.: Imagenet classification with deep convolutional neural networks. In: Neural Information Processing Systems, 2012.

[Ksi13] Ksibi A., Feki G., Ben Ammar A. & Ben Amar C., Effective Diversification for Ambiguous Queries in Social Image Retrieval, Computer Analysis of Images and Patterns, Springer Berlin Heidelberg, pp. 571-578, 2013.

[Lar13] Larsen J.E., Cuttone A., Jorgensen S.L., QS Spiral: Visualizing periodic quantified self-data, In proceedings of CHI 2013 Workshop on Personal Informatics in the Wild, 2013.

[Lee08] Lee H., Smeaton A.F., O'Connor N.E. et al., Constructing a SenseCam Visual Diary as a Media Process, Multimedia Systems 14: 341, 2008.

[Lid15] Lidon A., Bolanos M., Dimiccoli M., Radeva P., GaroleraM., Giro-i-Nieto X., Semantic summarization of egocentric photostream events, CoRR,http://arxiv.org/abs/1511.00438, 2015.

[Lin16] Lin H. L., Chiang T.-C., Chen L.P., Yang P.C. Image Searching by Events with Deep Learning for NTCIR-12 Lifelog. In The 12th NTCIR Conference, Tokyo, Japan, 2016.

[Mol16] del Molino A. G., Xu Q., Lim J.-H., Describing Lifelogs with Convolutional Neural Networks: A Comparative Study, LTA '16 Proceedings of the first Workshop on Lifelogging Tools and Applications, pp 39-44, 2016.

[Mol17] del Molino A. G., Mandal B., Lin J., Lim J., Subbaraju V. & Chandrasekhar V. VC-I2R@ImageCLEF2017: Ensemble of Deep Learned Features for Lifelog Video Summarization. Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017.

[Oli16a] Oliveira Barra G., Ayala A. C., Bolanos M., Dimiccoli M., Giro i Nieto X. & Radeva, P., LEMoRe: A Lifelog Engine for Moments Retrieval at the NTCIR-Lifelog LSAT Task, Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies, National Center of Sciences, Tokyo, Japan, June 7-10, 2016.

[Oli16b] Oliveira-Barra G., Dimiccoli M. & Radeva P., Egocentric Image Retrieval with Convolutional Neural Networks, Proceedings of the 19th International Conference of the Catalan Association for Artificial Intelligence, Barcelona, Catalonia, Spain, October 19-21, pp 71-76, 2016.

[Oli17] Oliveira-Barra G., Dimiccoli M, & Radeva P. Leveraging Activity Indexing for Egocentric Image Retrieval. In: Alexandre L., Salvador Sanchez J., Rodrigues J. (eds) Pattern Recognition and Image Analysis, IbPRIA, Lecture Notes in Computer Science, vol 10255. Springer, Cham, 2017.

[Pen17] Peng C., Zhang X., Yu G., Luo G., Sun J., Large Kernel Matters - Improve Semantic Segmentation by Global Convolutional Network, CoRR abs/1703.02719, 2017.

[Rey16] Reyes R. C. Time-sensitive Egocentric Image Retrieval for Findings Objects in Life-logs, Degree's Thesis Sciences and Telecommunication Technologies Engineering, Technical University of Catalonia, 2016.

[Saf16] Safadi B., Mulhem P., Quénot G., Chevallet J.P. LIG-MRIM at NTCIR-12 Lifelog Semantic Access Task, In The 12th NTCIR Conference, Tokyo, Japan, 2016.

[San17] Santoro A., Raposo D., Barrett D. G., Malinowski M., Pascanu R, Battaglia P., Lillicrap T., A simple neural network module for relational reasoning, arXiv preprint arXiv:1706.01427, 2017.

[Sel10] Sellen A. J. , Whittaker S., Beyond total capture: a constructive critique of lifelogging. Commun. ACM 53, pp 70-77, 2010.

[Wal10] Wali A.,Ben Aoun N., Karray H., Ben Amar C. & Alimi A. M., A New System for Event Detection from Video Surveillance Sequences Advanced Concepts for Intelligent Vision Systems, Springer Berlin Heidelberg, pp110-120, 2010.

[Xia16] Xia L., Ma Y., Fan W. VTIR at the NTCIR-12 2016 Lifelog Semantic Access Task. In The 12th NTCIR Conference, Evaluation of Information Access Technologies. Tokyo, Japan, 2016.

[Zhou17] Zhou L., Piras L., Riegler M., Boato G., Nguyen D. T. D., Gurrin C. Organizer Team at ImageCLEFlifelog 2017: Baseline Approaches for Lifelog Retrieval and Summarization, 2017.

 ISBN 978-80-86943-42-8